IBM SPSS Analytic Server
Versão 2

Guia do Usuário

Nota
Antes de utilizar estas informações e o produto suportado por elas, leia as informações em "Avisos" na página 35.

Informações sobre o Produto

Esta edição aplica-se à versão 2, liberação 0, modificação 0 de IBM SPSS Analytic Servere a todas as liberações e modificações subsequentes, até que seja indicado de outra forma em novas edições.

Índice

Capítulo 1. O que há de novo para os usuários na versão 2

Console do Analytic Server

Novo layout

O layout foi alterado, para que as páginas sejam acessadas por meio de uma página inicial, em vez de sanfonas.

Origens de dados

- É possível definir atributos customizados para a origem de dados e visualizar os atributos customizados criados pelos aplicativos.
- Ao criar os metadados para uma origem de dados, será possível iniciar uma varredura de todos os valores de dados para determinar os valores de categoria e os limites do intervalo. Varrer todos os valores de dados garante que os metadados estejam corretos, mas poderá levar algum tempo, se a origem de dados tiver muitos campos e registros.
- Há suporte para mais tipos de origens de dados.

  Tipo de conteúdo do arquivo

  O suporte para mais tipos de tipo de conteúdo do arquivo inclui definições adicionais e formatos de analisador. É possível também definir a ordem analisada de campos para cada arquivo em uma origem de dados. Ao incluir um diretório em uma origem de dados, será possível especificar regras para selecionar arquivos nesse diretório ou em seus subdiretórios.

  Arquivos semiestruturados

  Esses são arquivos, como logs da web, que não têm tanta estrutura quanto um arquivo de texto delimitado, mas contêm dados que podem ser extraídos em registros e campos por meio de expressões regulares.

  Arquivos compactados

  Os formatos de compactação suportados incluem Gzip, Deflate, Bz2, Snappy e IBM CMX. Além disso, os arquivos de sequência com qualquer um dos formatos de compactação mencionados anteriormente são suportados.

  Arquivos baseados em texto em formatos diferentes

  Uma origem de dados baseadas em texto simples agora podem conter documentos em diferentes formatos (PDF, Microsoft Word, etc.) para a analítica de texto.

  Arquivos do SPSS Statistics

  Os arquivos SPSS Statistics (*.sav, *.zsav) são arquivos binários que contêm um modelo de dados.

  Arquivos de formato binário divisível (*.asbf)

  Esse tipo de arquivo é, algumas vezes, produzido pelo Analytic Server; por exemplo, quando a análise requer o uso de campos que tem valores de lista.

  Arquivos de sequência

  Os arquivos de sequência (*.seq) são arquivos de texto estruturados como pares de chave/valor. Normalmente, eles são usados como um formato intermediário nas tarefas MapReduce.

  Tipo de conteúdo do banco de dados

  Será possível definir as origens de dados para Greenplum, MySQL e Sybase IQ, se o Analytic Server tiver sido configurado para ser capaz de usar essas origens de dados.

© Copyright IBM Corporation 2010, 2015

Tipo de conteúdo do HCatalog
É possível definir as origens de dados para Apache Cassandra, MongoDB e Oracle NoSQL, se o Analytic Server tiver sido configurado para ser capaz de usar essas origens de dados.

Tipo de conteúdo geo-espacial
É possível definir as origens de dados para as geografias usando os arquivos de formato ou os serviços de mapa online.

Analytics

Nova funcionalidade do SPSS Modeler

Mesclar
Suporte incluído para mesclar por condição de ranqueamento.

Série temporal
Suporte incluído para processamento de série temporal, além da construção distribuída e da pontuação de modelos casuais temporais (TCM). Consulte os nós Intervalos de Tempo AS, TCM de Fluxo e TCM no SPSS Modeler.

Dados espaciais
Suporte incluído para processamento de sistemas de coordenadas geográficas, além da construção distribuída e da pontuação de modelos de regras de associação geoespaciais (GSAR) e de processo de ponto de espaço-temporal (STP). Consulte os nós Reprojeção, Regras de Associação e STP no SPSS Modeler.

Armazenamento em cluster
Suporte incluído para construção distribuída e pontuação de modelos de cluster de duas etapas. Consulte o nó TwoStep-AS no SPSS Modeler.

Suporte aprimorado para a funcionalidade existente do SPSS Modeler

Agregado
Os campos de sequência podem ser agregados usando mín, máx, e contagem de valores não nulos. As estatísticas de ordem aproximada (mediana, quartil) são suportadas para campos numéricos na guia Otimização.

Mesclar
Suporte incluído para mesclar por condição e mescle por chaves sem chaves; por exemplo, para produzir uma média global.

Modelagem da combinação

O algoritmo para construir modelos de combinação para os modelos Árvore, Linear e Rede Neural é aprimorado para melhor lidar com os dados que não são distribuídos aleatoriamente em blocos dimensionados uniformemente.

Capítulo 2. Console do Analytic Server

O Analytic Server fornece uma interface de thin client para gerenciar origens de dados e projetos.

Efetuando login

1. Insira a URL do Analytic Server na barra de endereço do seu navegador. A URL pode ser obtida a partir do administrador do servidor.
2. Insira o nome de usuário com o qual efetuar logon no servidor.
3. Insira a senha associada ao nome de usuário especificado.

Após efetuar login, o início do Console será exibido.

Navegando no Console

- O cabeçalho exibe o nome do produto, o nome do usuário que efetuou login atualmente e o link para o sistema de ajuda. O nome do usuário que efetuou login atualmente é a cabeça de uma lista suspensa que inclui o link de logout.
- A área de conteúdo exibe as ações que podem ser tomadas no início do Console.

Uma origem de dados é uma coleção de registros, mais um modelo de dados, que definem um conjunto de dados para análise. A origem dos registros pode ser um arquivo (texto delimitado, texto de largura fixa, Excel) no HDFS, um banco de dados ou um HCatalog. O modelo de dados define todos os metadados (nomes de campo, armazenamento, nível de medição, entre outros) necessários para a análise dos dados. Os proprietários da origem de dados podem conceder ou restringir o acesso às origens de dados.

Listagem de origens de dados

A página principal Origens de Dados fornece uma lista de origens de dados dos quais o usuário atual é membro.

- Clique no nome de uma origem de dados para exibir seus detalhes e editar suas propriedades.
- Digite na área de procura para filtrar a listagem para exibir somente origens de dados com a sequência de procura em seu nome.
- Clique em **Novo** para criar uma nova origem de dados com o nome e o tipo de conteúdo especificados no diálogo **Incluir nova origem de dados**.
  - Consulte “Nomeando regras” na página 28 para obter restrições nos nomes que é possível fornecer às origens de dados.
  - Os tipos de conteúdo disponível são Arquivo, Banco de Dados, HCatalog e Geo-espacial.

    Nota: A opção HCatalog somente estará disponível, se o Analytic Server foi configurado para funcionar com essas origens de dados.

    Nota: O tipo de conteúdo não poderá ser editado após ser selecionado.
- Clique em **Excluir** para remover a origem de dados. Essa ação deixa todos os arquivos associados à origem de dados intactos.
- Clique em **Atualizar** para atualizar a listagem.
- A lista suspensa Ações executa a ação selecionada.
  1. Selecione **Exportar** para criar um archive da origem de dados e salvá-lo no sistema de arquivos local.

2. Selecione **Importar** para importar um archive criado pela ação Exportar.
3. Selecione **Duplicar** para criar uma cópia da origem de dados.

Detalhes da origem de dados individual

A área de conteúdo é dividida em diversas seções, que pode depender do tipo de conteúdo da origem de dados.

Detalhes

Essas configurações são comuns a todos os tipos de conteúdo.

Nome Um campo de texto editável que mostra o nome da origem de dados.

Nome de exibição

Um campo de texto editável que mostra o nome da origem de dados, conforme exibido em outros aplicativos. Se estiver em branco, o Nome será usado como o nome de exibição.

Descrição

Um campo de texto editável para fornecer um texto explicativo sobre a origem de dados.

É público

Uma caixa de seleção que indica se alguém pode ver a origem de dados (marcada) ou se os usuários ou grupos devem ser incluídos explicitamente como membros (desmarcada).

Atributos customizados

Os aplicativos podem anexar propriedades às origens de dados, como se a origem de dados fosse provisória, através do uso de atributos customizados. Esses atributos são expostos no console do Analytic Server para fornecer insight adicional em como os aplicativos usam a origem de dados.

Clique em **Salvar** para manter o estado atual das configurações.

Compartilhamento

Essas configurações são comuns a todos os tipos de conteúdo.

É possível compartilhar a propriedade de uma origem de dados incluindo usuários e grupos como autores.

- Digitando os filtros da caixa de texto nos usuários e grupos com a sequência de caracteres de procura em seu nome. Clique em **Incluir membro** para incluí-los na lista de autores.
- Para remover um autor, selecione um usuário ou grupo na lista de membros e clique em **Remover membro**.

Nota: Os administradores têm acesso de leitura e gravação a cada origem de dados, independentemente se eles estão especificamente listados como um membro.

Entrada de Arquivo

Configurações que são específicas à definição de origem de dados com o tipo de conteúdo de arquivo.

Visualizador de Arquivos

Mostra os arquivos disponíveis para inclusão na origem de dados. Selecione o modo **Projetos** para visualizar arquivos na estrutura do projeto do Analytic Server, **Origem de dados** para visualizar arquivos armazenados dentro de uma origem de dados ou **Sistema de arquivos** para visualizar o sistema de arquivos (normalmente HDFS). É possível pesquisar a estrutura da pasta, mas o HDFS não é editável para todos e no modo **Projetos**, não é possível incluir arquivos, criar pastas ou excluir itens no nível raiz, mas somente em projetos definidos. Para criar, editar ou excluir um projeto, use Projetos.

- Clique em **Fazer upload** para fazer upload de um arquivo para a origem de dados atuais ou projeto/subpasta. É possível navegar e selecionar diversos arquivos em um único diretório.
- Clique em **Nova pasta** para criar uma nova pasta na pasta atual, com o nome especificado no diálogo Novo nome da pasta.
- Clique em **Fazer download** para fazer download dos arquivos selecionados no sistema de arquivos local.
- Clique em **Excluir** para remover arquivos/pastas selecionadas.

Arquivos incluídos na definição da origem de dados
Use o botão Mover para incluir arquivos e pastas selecionados na origem de dados ou removê-los dela. Para cada arquivo ou pasta selecionada na origem de dados, clique em Configurações para definir as especificações para a leitura do arquivo.

Quando diversos arquivos forem incluídos em uma origem de dados, eles deverão compartilhar um metadado comum; isto é, cada arquivo deve ter o mesmo número de campos, os campos devem ser analisados na mesma ordem em cada arquivo e cada campo deve ter o mesmo armazenamento em todos os arquivos. Incompatibilidades entre os arquivos podem fazer com que o console falhe ao criar a Visualização e Metadados ou, caso contrário, os valores válidos serão analisados como inválidos (nulo), quando o Analytic Server ler o arquivo.

Seleções de Banco de Dados

Especifique parâmetros de conexão para o banco de dados que contém o conteúdo de registro.

Banco de dados
Selecione o tipo de banco de dados ao qual se conectar. Escolha a partir de: DB2, Greenplum, MySQL, Netezza, Oracle, SQL Server, Sybase IQ ou TeraData. Se o tipo que estiver sendo procurado não estiver listado, peça ao administrador do servidor para configurar o Analytic Server com o JDBC driver apropriado.

Endereço do servidor
Insira a URL do servidor que hospeda o banco de dados.

Porta do Servidor
O número da porta na qual o banco de dados recebe.

Nome do banco de dados
O nome do banco de dados ao qual você deseja se conectar.

Nome do Usuário
Se o banco de dados for protegido por senha, insira o nome de usuário.

Senha Se o banco de dados for protegido por senha, insira a senha.

Nome da tabela
Insira o nome de uma tabela do banco de dados que deseja usar.

Máximo de leituras simultâneas
Insira o limite no número de consultas paralelas que podem ser enviadas do Analytic Server para o banco de dados para ler a tabela especificada na origem de dados.

Seleções de HCatalog
Especifique os parâmetros para acessar os dados que são gerenciados no Apache HCatalog.

Banco de dados
O nome do banco de dados do HCatalog.

Nome da tabela
Insira o nome de uma tabela do banco de dados que deseja usar.

Filtro O filtro de partição para a tabela, se a tabela foi criada como tabela particionada. A filtragem de HCatalog é suportada somente nas chaves de partição de Hive da sequência de tipos.

Nota: Os operadores !=, <> e LIKE não parece funcionar em determinadas distribuições de Hadoop. Isso é um problema de compatibilidade entre HCatalog e as distribuições.

Mapeamentos de Campo do HCatalog
Exibe o mapeamento de um elemento no HCatalog para um campo na origem de dados. Clique em Editar para modificar os mapeamentos de campo.

Nota: Após criar uma origem de dados com base em HCatalog que expõe dados de uma tabela Hive, você poderá achar que quando a tabela Hive for formada por um número grande de arquivos de dados, ocorrerá um atraso substancial incorrido cada vez que o Analytic Server começar a ler os dados da origem de dados. Se você notar esse atraso, reconstrua a tabela Hive usando um número menor de arquivos de dados maiores e reduza o número de arquivos para 400 ou menos.

Seleções Geo-Espaciais
Especifique os parâmetros para acessar os dados geográficos.

Tipo Geo-Espacial
Os dados geográficos podem vir de um serviço de mapa online ou de um arquivo de forma.

Se você estiver usando um serviço de mapa, especifique a URL do serviço e selecione a camada de mapa que deseja usar.

Se você estiver usando um arquivo de forma, faça o upload do arquivo de forma.

Visualização e Metadados
Após especificar as configurações para a origem de dados, clique em Visualização e Metadados para verificar e confirmar as especificações da origem de dados.

Out As origens de dados com o tipo de conteúdo de banco de dados ou arquivo podem ser anexadas pela saída de fluxos que estão em execução no Analytic Server. Selecione **Tornar gravável** para ativar a anexação e:

- Para as origens de dados com o tipo de conteúdo do banco de dados, escolha uma tabela de base de dados de saída na qual os dados de saída são gravados.
- Para as origens de dados com o tipo de conteúdo de arquivos:
  1. Escolha uma pasta de saída na qual os novos arquivos são gravados.

     Dica: Use uma pasta separada para cada origem de dados, pois assim é mais fácil manter o controle das associações entre os arquivos e as origens de dados.
  2. Selecione um formato de arquivo: **CSV** (variável separada por vírgula) ou **Formato binário divisível**.
  3. Opcionalmente selecione **Tornar arquivo de sequência**. Isso será útil, se você desejar criar arquivos compactados divisíveis que são utilizáveis nas tarefas MapReduce de recebimento de dados.
  4. Selecione **As novas linhas podem ser escapadas** para fazer com que as novas linhas nos dados sejam gravadas como a sequência "\n" no arquivo de saída e a sequência "\n" seja gravada como "\\n" no arquivo de saída. Se desmarcada, a sequência de "\n" será gravada como "\n" no arquivo de saída e a presença de uma nova linha causará um erro.
  5. Selecione um formato de compactação. A lista inclui todos os formatos que foram configurados para uso com a instalação do Analytic Server.

Nota: Algumas combinações de formato de compactação e formato de arquivo resultam na saída que não pode ser dividida e são, portanto, inadequadas para o processamento MapReduce adicional. O Analytic Server produz um aviso na seção Saída ao fazer essa seleção.

Configurações (Origens de Dados do Arquivo)

O diálogo Configurações permite definir as especificações para dados baseados em arquivos de leitura. As configurações se aplicam a todos os arquivos selecionados e a todos os arquivos dentro das pastas selecionadas que correspondem aos critérios na guia **Seleção de arquivo**.

Especificar as configurações do analisador incorreto para um arquivo pode fazer com que o console falhe ao criar a Visualização e Metadados ou, caso contrário, os valores válidos serão analisados como inválidos (nulo), quando o Analytic Server ler o arquivo.

Guia Configurações

A guia Configurações permite especificar o tipo de arquivo e configurações de analisador específico para o tipo de arquivo.

É possível definir origens de dados usando arquivos compactados para qualquer formato de arquivo suportado. Os formatos de compactação suportados incluem Gzip, Deflate, Bz2, Snappy e IBM CMX.

Tipo de arquivo delimitado

Os arquivos delimitados são arquivos de texto de campo livre, cujos registros contêm um número constante de arquivos, mas um número variado de caracteres por campo. Geralmente, os arquivos delimitados têm as extensões de arquivo *.csv ou *.tab. Consulte “Configurações de tipo de arquivo delimitado” na página 8 para obter mais informações.

Tipo de arquivo fixo

Os arquivos de texto de campo fixo são arquivos cujos campos não são delimitados, mas iniciam na mesma posição e têm um comprimento fixo. Geralmente, os arquivos de texto de campo fixo possuem uma extensão de arquivo *.dat. Consulte “Configurações de tipo de arquivo fixo” na página 9 para obter mais informações.

Tipo de arquivo semiestruturado

Os arquivos semiestruturados (como *.log) são arquivos de texto que possuem uma estrutura previsível que pode ser mapeada para campos por meio de expressões regulares, mas não são tão altamente estruturados como arquivos delimitados. Consulte “Configurações de tipo de arquivo semiestruturado” na página 10 para obter mais informações.

Tipo de arquivo Text Analytics

Os arquivos Text Analytics são documentos (como *.doc, *.pdf ou *.txt) que podem ser analisados usando o SPSS Text Analytics.

Ignorar linhas vazias
Especifica se deve ignorar as linhas vazias no conteúdo de texto extraído. O padrão é **Não**.

Separador de linha
Especifica a sequência que define uma nova linha. O padrão é o caractere de nova linha "\n".

Tipo de SPSS Statistics

Os arquivos SPSS Statistics (*.sav, *.zsav) são arquivos binários que contêm um modelo de dados. Não é necessária nenhuma configuração adicional na guia Configurações para este tipo de arquivo.

Tipo de arquivo de formato binário divisível

Especifica se o tipo de arquivo é um arquivo de formato binário divisível (*.asbf). Esse tipo de arquivo é, algumas vezes, produzido pelo Analytic Server; por exemplo, quando a análise requer o uso de campos que tem valores de lista. Não é necessária nenhuma configuração adicional na guia Configurações para este tipo de arquivo.

Tipo de arquivo de sequência

Os arquivos de sequência (*.seq) são arquivos de texto estruturados como pares de chave/valor. Normalmente, eles são usados como um formato intermediário nas tarefas MapReduce.

Tipo de arquivo Excel

Especifica se o tipo de arquivo é um arquivo do Microsoft Excel (*.xls, *.xlsx). Consulte “Configurações de tipo de arquivo do Excel” na página 11 para obter mais informações.

Configurações de tipo de arquivo delimitado:

É possível especificar as configurações a seguir para os tipos de arquivos delimitados.

Codificação de Conjunto de Caracteres
A codificação de caracteres do arquivo. Selecione ou especifique o nome do conjunto de caracteres Java como "UTF-8", "ISO-8859-2", "GB18030". O padrão é **UTF-8**.

Delimitadores de campo
Um ou mais caracteres marcando os limites de campo. Cada caractere é obtido como um delimitador independente. Por exemplo, se você selecionar **Vírgula** e **Tabulação** (ou selecionar **Outro** e digitar `,\t`), significa que uma tabulação marca os limites do campo. Se os caracteres de controle delimitam campos, os caracteres especificados aqui são tratados como delimitadores, além de caracteres de controle. O padrão será "," se os caracteres de controle não delimitarem os campos; caso contrário, o padrão será uma sequência de caracteres vazia.

Caracteres de controle delimitam campos
Configura se os caracteres de controle ASCII, exceto LF e CR, são tratados como delimitadores de campo. Assume como padrão **Não**.

A primeira linha contém nomes de campo
Configura se a primeira linha será usada para determinar os nomes de campo. Assume como padrão **Não**.

Número de caracteres iniciais a ignorar
O número de caracteres no início do arquivo a ser ignorado. Um número inteiro não negativo. O padrão é 0.

Mesclar espaço em branco
Configura se diversas ocorrências adjacentes de espaço e/ou tabulação serão tratadas como um delimitador de campo único. Não terá efeito se nem um espaço e nem uma tabulação for um delimitador de campo. O padrão é **Sim**.

Caracteres de comentário de final de linha
Um ou mais caracteres que marcam os comentários de final de linha. O caractere e tudo depois dele no registro são ignorados. Cada caractere é obtido como um marcador de comentário independente. Por exemplo, "/*" significa que uma barra ou um asterisco inicia um comentário. Não é possível definir marcadores de comentários multicaracteres, como "//". A sequência vazia sinaliza que nenhum caractere de comentário está definido. Se estiver definida, os caracteres de comentário serão verificados antes que as aspas sejam processadas ou que os caracteres iniciais a serem ignorados sejam ignorados. O padrão é a sequência vazia.

Caracteres inválidos
Determina como caracteres inválidos (sequências de bytes que não correspondem aos caracteres na codificação) devem ser manipulados. Uma sequência de caracteres vazia indica que eles devem ser descartados. Uma sequência de caracteres não vazia (geralmente um único caractere) indica que eles devem ser substituídos pelo conteúdo da sequência. O padrão é a sequência vazia.

Aspas simples
Especifica a manipulação de aspas simples (apóstrofos). O padrão é **Manter**.

Manter
Aspas simples não têm nenhum significado especial e são tratadas como qualquer outro caractere.

Drop Aspas simples são excluídas, a menos que entre aspas

Par Aspas simples são tratadas como caracteres aspas simples e caracteres entre pares de aspas simples perdem qualquer significado especial (eles são considerados entre aspas). Se as próprias aspas simples podem ocorrer dentro de sequências entre aspas é determinado por meio da configuração de **Aspas podem ser colocadas entre aspas por meio da duplicação**.

Aspas duplas
Especifica a manipulação de aspas duplas. O padrão é **Par**.

Manter
As aspas duplas não têm significado especial e são tratadas como qualquer outro caractere.

Drop As aspas duplas são excluídas, a menos que entre aspas

Par As aspas duplas são tratadas como caracteres de aspas e os caracteres entre pares de aspas duplas perdem qualquer significado especial (eles são considerados entre aspas). Se as aspas duplas em si puderem ocorrer dentro de sequências entre aspas duplas, será determinado pela configuração **Aspas podem ser colocadas entre aspas por duplicação**.

Aspas podem ser colocadas entre aspas por meio de duplicação
Indica se as aspas duplas poderão ser representadas em sequências entre aspas duplas e as aspas simples poderão ser representadas em sequências entre aspas simples, quando configuradas como **Par**. Se **Sim**, as aspas duplas serão escapadas dentro de sequências entre aspas duplas por duplicação e as aspas simples serão escapadas dentro de sequências entre aspas simples por duplicação. Se **Não**, não há maneira de colocar aspas duplas entre aspas dentro de uma sequência entre aspas duplas ou aspas simples dentro de uma sequência entre aspas simples. O padrão é **Sim**.

As novas linhas podem ser escapadas
Indica se o analisador interpreta a sequência de caracteres "\n" como uma nova linha ao ler um arquivo. Se as novas linhas não forem escapadas, "\n" simplesmente será lido como uma sequência de caracteres. Se as novas linhas forem escapadas, "\n" será lido como caractere de nova linha ASCII e "\\n" será lido como a sequência de caracteres "\n". O padrão é **Não**.

Configurações de tipo de arquivo fixo:

É possível especificar as configurações a seguir para os tipos de arquivos fixos.

Codificação de Conjunto de Caracteres
A codificação de caracteres do arquivo. Selecione ou especifique o nome do conjunto de caracteres Java como "UTF-8", "ISO-8859-2", "GB18030". O padrão é **UTF-8**.

Caracteres inválidos
Determina como caracteres inválidos (sequências de bytes que não correspondem aos caracteres na codificação) devem ser manipulados. Uma sequência de caracteres vazia indica que eles

devem ser descartados. Uma sequência de caracteres não vazia (geralmente um único caractere) indica que eles devem ser substituídos pelo conteúdo da sequência. O padrão é a sequência vazia.

Comprimento do registro
Indica quantos registros estão definidos. Se **Nova linha delimitada**, os registros serão definidos (delimitados) por novas linhas, iniciando do arquivo ou o término do arquivo. Se **Duração específica**, os registros serão definidos pela duração do registro em bytes. Especifique um valor positivo.

Registros iniciais a serem ignorados
O número de registros no início do arquivo a ser ignorado. Especifique um número inteiro não negativo. O valor padrão é 0.

Campos
Essa seção define os campos no arquivo. Clique em **Incluir campo** e especifique o nome do campo, a coluna na qual os valores do campo iniciam e o comprimento dos valores do campo. As colunas em um arquivo são numeradas começando em 0.

Configurações de tipo de arquivo semiestruturado:

As configurações para arquivos semiestruturados consistem em regras para mapeamento do conteúdo do arquivo para campos.

Tabela de regras
As regras individuais extraem informações de um registro para criar um campo; juntas na tabela de regras, elas definem todos os campos que podem ser extraídos de cada registro em uma origem de dados.

As regras na tabela são aplicadas na ordem a cada registro; se todas as regras na tabela corresponderem ao registro, as outras tabelas de regras não serão necessárias para processar o registro e o próximo registro será processado. Se nenhuma regra na tabela não corresponder, todos os valores dos campos extraídos por regras anteriores na tabela serão descartados; se houver outras tabelas de regras, as regras nessa tabela serão aplicadas ao registro. Se nenhuma tabela corresponder ao registro, a regra Incompatibilidade será aplicada.

Incompatibilidade
É possível escolher **Ignorar** os registros que não correspondem a nenhuma das tabelas de regras ou configurar o valor de todos os campos no registro como **Ausente** (nulo).

Exportar regras
É possível salvar a tabela de regras visíveis atualmente para reutilização. A tabela exportada é salva no servidor.

Importar regras
É possível importar uma tabela de regras salvas na tabela de regras atualmente visível. Isso sobrescreve as regras definidas para essa tabela, portanto, é melhor criar uma nova tabela e, em seguida, importar uma tabela de regras.

Editor de regras

O editor de regras permite criar uma regra de extração para um único campo.

Grupo de captura anônima
Uma regra de captura de campo, normalmente é iniciada para extrair dados de um registro na posição em que a regra anterior foi interrompida. Quando houver informações estranhas entre dois campos em uma origem de dados semiestruturados, elas poderão ser úteis para definir um grupo de captura anônima que posiciona o analisador onde o próximo campo é iniciado. Ao selecionar **Grupo de captura anônima**, os controles para nomear e identificar o grupo de captura serão desativados, mas o restante do diálogo funcionará normalmente.

Nome do campo
Insira um nome para o campo. Isso é usado para definir os metadados da origem de dados. Os nomes do campo devem ser exclusivos em uma tabela de regras.

Nome da regra
Opcionalmente, digite uma etiqueta descritiva para a regra.

Descrição
Opcionalmente, digite uma descrição mais longa para a regra.

Definindo uma regra
Há dois métodos para definir as regras.

Use os controles de regras de extração
Isto simplifica a criação de regras de extração.

1. Especifique o ponto para iniciar a extração de dados do campo; **Posição atual** começará onde a regra anterior foi interrompida e **Ignorar até** começará no início do registro e ignorará todos os caracteres até que ele atinja um ponto especificado na caixa de texto. Selecione **Incluir**, se desejar que os dados do campo incluam o caractere na posição inicial.
2. Selecione um grupo de captura de campo na lista suspensa **Capturar**.
3. Opcionalmente, selecione o ponto para parar a extração de dados do campo; **Espaços em branco** parará quando os caracteres de espaço em branco (como espaços ou tabulações) forem encontrados e **Em caractere(s)** parará na sequência de caracteres especificada. Selecione **Incluir**, se desejar que os dados do campo incluam o caractere na posição de parada.

Definir regras regexp manualmente
Selecione essa opção, se estiver confortável gravando a sintaxe de expressão regular. Insira uma expressão regular na caixa de texto **Regexp**.

Incluir grupo de captura de campo
Isto permite salvar a expressão regular para uso posterior. O grupo de captura salva aparece na lista suspensa **Capturar**.

O Editor de Regras mostrará uma visualização dos dados extraídos do primeiro registro por essa regra, depois que todas as regras anteriores na tabela de regras foram aplicadas.

Configurações de tipo de arquivo do Excel:

É possível especificar as configurações a seguir para arquivos do Excel.

Seleção de planilha
Seleciona a planilha do Excel a ser usada como a origem de dados. Especifique um índice numérico (o índice da primeira planilha é 0) ou o nome da planilha. O padrão é usar a primeira planilha.

Seleção de intervalo de dados para importação.
É possível importar dados que iniciam com a primeira linha sem espaços em branco ou com um intervalo explícito de células.

- **O intervalo inicia na primeira linha sem espaços em branco**. Localiza a primeira célula sem espaços em branco e usa isso como o canto superior esquerdo do intervalo de dados.
- Como alternativa, especifique um intervalo explícito de células por linha e coluna. Por exemplo, para especificar o intervalo do Excel A1:D5, é possível inserir A1 no primeiro campo e D5 no segundo (ou, como alternativa, R1C1 e R5C4). Todas as linhas no intervalo especificado são retornadas, incluindo as linhas em branco.

A primeira linha contém nomes de campo
Especifica se a primeira linha do intervalo de células selecionadas conterá os nomes do campo. O padrão é **Não**.

Pare a leitura após encontrar linhas em branco
Especifica se deve parar a leitura de registros após encontrar mais de uma linha em branco ou, se deve continuar a leitura de todos os dados até o final da planilha, incluindo as linhas em branco. O padrão é **Não**.

Formatos

A guia Formatos permite definir informações de formatação para os campos analisados.

Configurações de conversão de campo

Cortar espaço em branco
Remove os caracteres de espaço em branco do início e/ou do final dos campos de sequência. Padroniza como **Nenhum**. Os seguintes valores são suportados:

Nenhuma
Não remova caracteres de espaço em branco.

Esquerdo
Remove caracteres de espaço em branco do início da sequência.

Direito
Remove caracteres de espaço em branco do final da sequência.

Ambas
Remove caracteres de espaço em branco do início e do final da sequência.

Código do idioma
Define um código de idioma. Assume como padrão o código de idioma do servidor. A sequência do código de idioma deve ser especifica como: <language>[_country[_variant]], em que:

idioma
Um código válido de duas letras minúsculas, conforme definido pela ISO-639.

país Um código válido de duas letras maiúsculas, conforme definido pela ISO-3166.

variante
Um código de fornecedor ou específico do navegador.

Separador decimal
Configura o caractere usado como o sinal decimal. Padroniza como a configuração específica ao código de idioma.

Símbolos de agrupamento
Define se o caractere específico de código de idioma usado para o separador de milhares deve ou não ser usado.

Formato de data padrão
Define um formato de data padrão. Todos os padrões de formatos definidos pela especificação Locale data markup Language (LDML) Unicode são suportados.

Formato de horário padrão
Define um formato de horário padrão.

Registro de data e hora padrão
Define um formato de registro de data e hora padrão.

Fuso horário padrão
Configura o fuso horário. Padroniza como UTC. A configuração aplica-se aos campos de horário e do registro de data e hora que não têm fuso horário especificado explicitamente.

Substitutos de campo

Esta seção permite designar as instruções de formatação aos campos individuais. Selecione um campo no modelo de dados ou digite em um nome do campo e clique em **Incluir** para incluí-lo na lista de campos com instruções individuais. Clique em **Remover** para removê-lo dessa lista. Para um campo selecionado na lista, é possível configurar as seguintes propriedades do campo.

Armazenamento
Configure o armazenamento do campo.

Separador decimal
Para os campos com armazenamento real, defina o caractere usado como o sinal decimal. Padroniza como a configuração específica ao código de idioma.

Símbolos de agrupamento
Para os campos com armazenamento Inteiro ou Real, configura se o caractere específico ao código de idioma usado para o separador de milhares deve ou não ser usado.

Formatos
Para os campos com armazenamento de Data, Hora ou Registro de data e hora, defina o formato. Escolha um formato na lista suspensa.

Guia Ordem do campo

Para os tipos de arquivos delimitados e Excel, a guia Ordem do campo permite definir a ordem analisada dos campos para o arquivo. Isso será importante quando houver diversos arquivos em uma origem de dados, porque a ordem real dos campos pode ser diferente nos arquivos, mas a ordem analisada dos campos deve ser a mesma para criar um modelo de dados consistentes.

Para os tipos de arquivo fixo e semiestruturado, a ordem é definida na guia Configurações.

Quando houver um único arquivo na origem de dados, ou todos os arquivos tiverem a mesma ordem de campo, será possível usar o padrão **A ordem do campo corresponde ao modelo de dados**. Se houver diversos arquivos na origem de dados e a ordem dos campos no arquivos não corresponder, defina uma **Ordem específica do campo** para a análise do arquivo.

1. Para incluir um campo na lista ordenada, digite o nome do campo ou selecione-o na lista fornecida pelo modelo de dados. É possível incluir todos os campos no modelo de dados de uma vez, clicando em **Incluir tudo**. Os nomes do campo somente serão incluídos uma vez na lista ordenada.
2. Use os botões de seta para ordenar os campos, conforme desejado.

Quando a **Ordem específica do campo** for usada, os campos não incluídos na lista não farão parte do conjunto de resultados para esse arquivo. Se houver campos que estiverem no modelo de dados que não estiverem listados nesse diálogo, os valores serão nulo no conjunto de resultados.

Guia Pasta

Ao especificar as configurações do analisador para uma pasta, a guia Pasta permitirá escolher quais arquivos na pasta serão incluídos na origem de dados.

Corresponder todos os arquivos na pasta selecionada
A origem de dados inclui todos os arquivos no nível superior da pasta; os arquivos nas subpastas não são incluídos.

Corresponder arquivos usando uma expressão regular
A origem de dados inclui todos os arquivos no nível superior da pasta que corresponde à expressão regular especificada; os arquivos nas subpastas não são incluídos.

Corresponder arquivos usando uma expressão Unix globbing (potencialmente recursiva)
A origem de dados inclui todos os arquivos que correspondem à expressão Unix globbing especificada; a expressão pode incluir arquivos que estão nas subpastas da pasta selecionada.

Mapeamentos de Campo do HCatalog

Esquema do HCatalog

Exibe a estrutura da tabela especificada. O HCatalog pode suportar um conjunto de dados altamente estruturados. Para definir uma origem de dados do Analytic Server nesses dados, a estrutura deverá ser simplificada em linhas e colunas simples. Selecione um elemento no esquema e clique no botão Mover para mapeá-lo para um campo para análise.

Nem todos os nós de árvore podem ser mapeados. Por exemplo, uma matriz ou um mapa de tipos complexos é considerado um "pai" e não pode ser mapeado diretamente; cada elemento simples em uma matriz ou mapa de HCatalog deve ser incluído separadamente. Esses nós podem ser identificados pelo rótulo no final da árvore em ...:array:struct ou ...:map:struct.

Por exemplo:

- Para uma matriz de números inteiros, é possível designar um campo a um valor dentro da matriz: bigintarray[45], mas não a própria matriz: bigintarray
- Para um mapa, é possível designar um campo a um valor dentro do mapa: datamap["key"], mas não o próprio mapa: datamap
- Para uma matriz de uma matriz de números inteiros, é possível designar um campo a um valor bigintarrayarray[45][2], mas não a própria matriz, bigintarrayarray[45].

  Portanto, ao designar um campo a uma matriz ou elemento de mapa, a definição do elemento deve incluir o índice ou a chave: bigintarray[index] ou bigintmap["key"].

Mapeamentos de campo

Elemento do HCatalog

Clique duas vezes em uma célula para editar. Você deverá editar a célula quando o elemento do HCatalog for uma matriz ou mapa. Com uma matriz, especifique o número inteiro que corresponda ao membro da matriz que você deseja mapear para um campo. Com um mapa, especifique uma sequência de caracteres entre aspas que corresponda à chave que você deseja mapear para um campo.

Campo de Mapeamento

O campo da maneira em que aparece na origem de dados do Analytic Server. Clique duas vezes em uma célula para editar. Os valores duplicados na coluna Campo de Mapeamento não são permitidos e resultam em um erro.

Armazenamento

O armazenamento do campo. O armazenamento é derivado do HCatalog e não pode ser editado.

Nota: Quando Visualização e Metadados é clicado para finalizar uma origem de dados do HCatalog, não aparecem opções de edição.

Dados brutos

Exibe os registros como eles estão armazenados no HCatalog; isso pode ajudar a determinar como mapear o esquema HCatalog para os campos.

Nota: Qualquer filtragem especificada no HCatalog Selections é aplicada à visualização dos dados brutos.

Ativando as origens de dados do HCatalog

O Analytic Server fornece suporte para origens de dados do HCatalog. Essa seção descreve como ativar vários bancos de dados do NoSQL subjacentes.

Apache Accumulo

O Analytic Server fornece suporte para origens de dados HCatalog que possuem um conteúdo subjacente no Apache Accumulo.

O armazenamento de chave/valor distribuído do Apache Accumulo é um armazenamento de dados e um sistema de recuperação, com base no design do BigTable do Google e é construído na parte superior do Apache Hadoop, do Zookeeper e do Thrift. O Apache Accumulo apresenta algumas novas melhorias no design do BigTable na forma de controle de acesso baseado em célula e em um mecanismo de programação do lado do servidor que pode modificar os pares de chave/valor em vários pontos no processo de gerenciamento de dados.

Para criar uma tabela externa do Apache Accumulo no Hive use a sintaxe a seguir:

```
set accumulo.instance.id=<instance_name>;
set accumulo.user.name=<user_name>;
set accumulo.user.pass=<user_password>;
set accumulo.zookeepers=<zookeeper_host_port>;

CREATE EXTERNAL TABLE <hive_table_name>(<table_column_specifications>)
STORED BY 'com.ibm.spss.hcatalog.AccumuloStorageHandler'
WITH SERDEPROPERTIES (
'accumulo.columns.mapping' = '<family_and_qualifier_mappings>',
'accumulo.table.name' = '<Accumulo_table_name>')
TBLPROPERTIES (
    "accumulo.instance.id"="<instance_name>",
    "accumulo.zookeepers"="<zookeeper_host_port>"
);
```

Por exemplo:

```
set accumulo.instance.id=<id>;
set accumulo.user.name=admin;
set accumulo.user.pass=test;
set accumulo.zookeepers=<host>:<port>;

CREATE EXTERNAL TABLE acc_drug1n(rowid STRING,age STRING,sex STRING,bp STRING,
    cholesterol STRING,na STRING,k STRING,drug STRING)
STORED BY 'com.ibm.spss.hcatalog.AccumuloStorageHandler'
WITH SERDEPROPERTIES (
'accumulo.columns.mapping' = 'rowID,drug|age,drug|sex,drug|bp,drug|cholesterol,
    drug|na,drug|k,drug|drug',
'accumulo.table.name' = 'drug1n')
TBLPROPERTIES (
    "accumulo.instance.id"="<id>",
    "accumulo.zookeepers"="<host>:<port>"
);
```

Nota: O nome de usuário e senha do Accumulo para a tabela fornecida do Accumulo deve corresponder ao nome de usuário e senha do usuário do Analytic Server autenticado.

Apache Cassandra

O Analytic Server fornece suporte para origens de dados HCatalog que possuem um conteúdo subjacente no Apache Cassandra.

O Cassandra fornece um armazenamento de chave-valor estruturado. Mapa de chaves para diversos valores, que são agrupados nas famílias de coluna. As famílias de coluna serão corrigidas quando um banco de dados for criado, mas as colunas podem ser incluídas em uma família a qualquer momento. Além disso, as colunas são incluídas somente para especificar as chaves, portanto, chaves diferentes podem ter números diferentes de colunas em qualquer família fornecida. Os valores de uma família de coluna para cada chave são armazenados juntos.

Há duas maneiras de definir as tabelas Cassandra: usando a interface da linha de comandos Cassandra (cassandra-cli) de legado e o novo shell CQL (csqlsh).

Use a sintaxe a seguir para criar uma tabela externa do Apache Cassandra no Hive, se a tabela tiver sido criada usando a CLI de legado.

```
CREATE EXTERNAL TABLE <hive_table_name>  (<column specifications>)
STORED BY 'com.ibm.spss.hcatalog.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<cassandra_column_family>",
"cassandra.host"="<cassandra_host>","cassandra.port" = "<cassandra_port>")
TBLPROPERTIES ("cassandra.ks.name" = "<cassandra_keyspace>");
```

Por exemplo, para a definição de tabela CLI a seguir:

```
create keyspace test
with placement_strategy = 'org.apache.cassandra.locator.SimpleStrategy'
and strategy_options = [{replication_factor:1}];

create column family users with comparator = UTF8Type;

update column family users with
        column_metadata =
        [
        {column_name: first, validation_class: UTF8Type},
        {column_name: last, validation_class: UTF8Type},
        {column_name: age, validation_class: UTF8Type, index_type: KEYS}
        ];

assume users keys as utf8;

set users['jsmith']['first'] = 'John';
set users['jsmith']['last'] = 'Smith';
set users['jsmith']['age'] = '38';
set users['jdoe']['first'] = 'John';
set users['jdoe']['last'] = 'Dow';
set users['jdoe']['age'] = '42';

get users['jdoe'];
```

... o DDL da tabela Hive será semelhante a este:

```
CREATE EXTERNAL TABLE cassandra_users  (key string, first string, last string, age string)
STORED BY 'com.ibm.spss.hcatalog.CassandraStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "users",
"cassandra.host"="<cassandra_host>","cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

Use a sintaxe a seguir para criar uma tabela externa do Apache Cassandra no Hive, se a tabela tiver sido criada usando o CQL.

```
CREATE EXTERNAL TABLE <hive_table_name>  (<column specifications>)
STORED BY 'com.ibm.spss.hcatalog.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "<cassandra_column_family>",
"cassandra.host"="<cassandra_host>","cassandra.port" = "<cassandra_port>")
TBLPROPERTIES ("cassandra.ks.name" = "<cassandra_keyspace>");
```

Por exemplo, para a definição de tabela CQL3 a seguir:

```
CREATE KEYSPACE TEST WITH REPLICATION = { 'class' : 'SimpleStrategy', 'replication_factor' : 2 };
USE TEST;

CREATE TABLE bankloan_10(
  row int,
  age int,
  ed int,
  employ int,
  address int,
  income int,
  debtinc double,
  creddebt double,
  othdebt double,
  default int,
  PRIMARY KEY(row)
);
```

```
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (1,41,3,17,12,176,9.3,11.359392,5.008608,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (2,27,1,10,6,31,17.3,1.362202,4.000798,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (3,40,1,15,14,55,5.5,0.856075,2.168925,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (4,41,1,15,14,120,2.9,2.65872,0.82128,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (5,24,2,2,0,28,17.3,1.787436,3.056564,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (6,41,2,5,5,25,10.2,0.3927,2.1573,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (7,39,1,20,9,67,30.6,3.833874,16.668126,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (8,43,1,12,11,38,3.6,0.128592,1.239408,0);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (9,24,1,3,4,19,24.4,1.358348,3.277652,1);
INSERT INTO bankloan_10 (row, age,ed,employ,address,income,debtinc,creddebt,othdebt,default)
   VALUES (10,36,1,0,13,25,19.7,2.7777,2.1473,0);
```

... o DDL da tabela Hive é o seguinte:

```
CREATE EXTERNAL TABLE cassandra_bankloan_10 (row int, age int,ed int,employ int,address int,
   income int,debtinc double,creddebt double,othdebt double,default int)
STORED BY 'com.ibm.spss.hcatalog.CassandraCqlStorageHandler'
WITH SERDEPROPERTIES("cassandra.cf.name" = "bankloan_10","cassandra.host"="<cassandra_host>",
   "cassandra.port" = "9160")
TBLPROPERTIES ("cassandra.ks.name" = "test");
```

Apache HBase

O Analytic Server fornece suporte para origens de dados HCatalog que possuem um conteúdo subjacente no Apache HBase.

O Apache HBase é um software livre, distribuído, com versão, de armazenamento orientado a coluna na parte superior do Hadoop e do HDFS.

Para criar uma tabela externa do HBase no Hive use a sintaxe a seguir:

```
CREATE EXTERNAL TABLE <tablename>(<table_column_specifications>)
STORED BY 'com.ibm.spss.hcatalog.HBaseStorageHandler'
WITH SERDEPROPERTIES ("hbase.columns.mapping" = "<column_mapping_spec>")
TBLPROPERTIES("hbase.table.name" = "<hbase_table_name>")
```

Por exemplo:

```
CREATE EXTERNAL TABLE hbase_drug1n(rowid STRING,age STRING,sex STRING,bp STRING,
    cholesterol STRING,na STRING,k STRING,drug STRING)
STORED BY 'com.ibm.spss.hcatalog.HBaseStorageHandler'
WITH SERDEPROPERTIES ("hbase.columns.mapping" = ":key,drug:age,drug:sex,drug:bp,
    drug:cholesterol,drug:na,drug:k,drug:drug")
TBLPROPERTIES("hbase.table.name" = "drug1n");
```

Nota: Para obter informações sobre como criar uma tabela HBase, consulte o Guia de referência do Apache HBase (http://hbase.apache.org/book.html).

Nota: É uma boa prática para prefaciar o nome do banco de dados para indicar o tipo de banco de dados. Por exemplo, nomeie um banco de dados HB_drug1n para indicar um banco de dados HBase ou um ACC_drug1n para indicar um banco de dados Accumulo. Isso ajudará com a seleção do arquivo HCatalog, quando no console do Analytic Server.

MongoDB

O Analytic Server fornece suporte para origens de dados HCatalog que possuem conteúdo subjacente no MongoDB.

O MongoDB é um banco de dados de documento de software livre e o banco de dados NoSQL principal gravado no C++. O banco de dados armazena documentos de estilo JSON com esquemas dinâmicos.

Para criar uma tabela externa MongoDB no Hive use a sintaxe a seguir:

```
create external table <hive_table_name>(<column specifications>)
stored by "com.ibm.spss.hcatalog.MongoDBStorageHandler"
with serdeproperties ( "mongo.column.mapping" = "<MongoDB to Hive mapping>" )
tblproperties ( "mongo.uri" = "'mongodb://<host>:<port>/<database>.<collection>" );
```

Por exemplo:

```
create external table mongo_bankloan(age bigint,ed bigint,employ bigint, address bigint,income bigint,
    debtinc double, creddebt double,othdebt double,default bigint)
STORED BY 'com.ibm.spss.hcatalog.MongoDBStorageHandler'
with serdeproperties ( 'mongo.column.mapping' = '{"age":"age","ed":"ed","employ":"employ","address":"address",
    "income":"income","debtinc":"debtinc","creddebt":"creddebt","othdebt":"othdebt","default":"default"}' )
tblproperties ('mongo.uri'='mongodb://9.48.11.162:27017/test.bankloan');
```

Oracle NoSQL

O Analytic Server fornece suporte para origens de dados HCatalog que possuem um conteúdo subjacente no Oracle NoSQL.

O Banco de dados Oracle NoSQL é um banco de dados de valor-chave distribuído. Os dados são armazenados como pares de valores de chave, que são gravados em nós de armazenamentos específicos, com base no valor do hash da chave principal. Os nós de armazenamento são replicados para garantir alta disponibilidade. Os aplicativos clientes são gravados usando a API Java/C API para ler e gravar dados.

SerDe e parâmetros da tabela

O manipulador de armazenamento do Oracle NoSQL suporta os parâmetros a seguir.

Parâmetros SERDEPROPERTIES

kv.major.keys.mapping
Lista separada por vírgula das chaves maiores. Obrigatório

kv.minor.keys.mapping
Lista separada por vírgula das chaves menores. Opcional

kv.parent.key
Especifica a chave-pai cujos pares de valores de chave "filho" devem ser retornados pela consulta. O caminho de chave maior deve ser um caminho parcial e o caminho de chave menor deve estar vazio. Opcional.

kv.avro.json.key
O nome da chave menor usada para conter o valor definido com o esquema Avro. Se a chave menor não estiver definida, o que geralmente é o caso, padronize como "valor". Se o parâmetro não estiver definido, o valor será retornado como uma sequência JSON. Opcional.

kv.avro.json.keys.mapping.column
Define o nome da coluna Hive para os pares de valores de chave maior/menor. A coluna Hive deve ter o tipo map<string,string>. Opcional.

Parâmetros TABLEPROPERTIES

kv.host.port
O endereço IP e o número da porta do banco de dados Oracle NoSQL. Obrigatório

kv.name
O nome do armazenamento de chave-valor do Oracle NoSQL. Requerido.

Exemplo: esquema Avro simples

O layout de dados é modelado usando a estrutura de serialização do Apache Avro. Para seguir essa abordagem, crie um esquema Avro; por exemplo:

```
{"type":"record",
 "name":"DrugSchema",
 "namespace":"avro",
 "fields":[
        {"name":"id","type":"string","default":""},
        {"name":"age","type":"string","default":""},
        {"name":"sex","type":"string","default":""},
        {"name":"bp","type":"string","default":""},
        {"name":"drug","type":"string","default":""}
]}
```

Esse esquema deve ser registrado com o banco de dados Oracle NoSQL e os dados preenchidos devem incluir uma referência ao esquema, conforme mostrado a seguir.

```
put -key /drugstore_avro/1 -value
   "{\"id\":\"1\",\"age\":\"23\",\"sex\":\"F\",\"bp\":\"HIGH\",\"drug\":\"drugY\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/2 -value
   "{\"id\":\"2\",\"age\":\"47\",\"sex\":\"M\",\"bp\":\"LOW\",\"drug\":\"drugC\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/3 -value
   "{\"id\":\"3\",\"age\":\"47\",\"sex\":\"M\",\"bp\":\"LOW\",\"drug\":\"drugC\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/4 -value
   "{\"id\":\"4\",\"age\":\"28\",\"sex\":\"F\",\"bp\":\"NORMAL\",\"drug\":\"drugX\"}"
   -json avro.DrugSchema
put -key /drugstore_avro/5 -value
   "{\"id\":\"5\",\"age\":\"61\",\"sex\":\"F\",\"bp\":\"LOW\",\"drug\":\"drugY\"}"
   -json avro.DrugSchema
```

Para expor os dados no Hive, crie uma tabela externa e especifique a propriedade adicional **kv.avro.json.key** na seção SERDEPROPERTIES. O valor da propriedade deve ser o nome da chave menor ou o nome predefinido **value**, se a chave menor não estiver definida.

```
CREATE EXTERNAL TABLE oracle_json(id string, age string, sex string, bp string, drug string)
  STORED BY 'com.ibm.spss.hcatalog.OracleKVStorageHandler'
  WITH SERDEPROPERTIES ("kv.major.keys.mapping" = "drugstore_avro,keyid",
        "kv.parent.key"="/drugstore_avro","kv.avro.json.key" = "value")
  TBLPROPERTIES ("kv.host.port" = "<hostname>:5000", "kv.name" = "kvstore");
```

Executar select * from oracle_json produz os resultados a seguir.

```
select * from oracle_json;

1 23 F HIGH drugY
5 61 F LOW drugY
3 47 M LOW drugC
2 47 M LOW drugC
4 28 F NORMAL drugX
```

A tabela oracle_json pode ser usada no console do Analytic Server para criar uma origem de dados do Oracle NoSQL.

Exemplo: chaves complexas

Agora considere o esquema Avro a seguir.

```
{ "type": "record",
  "name": "DrugSchema",
  "namespace":"avro",
  "fields": [
      {"name": "age", "type": "string", "default": ""}, // age
      {"name": "bp", "type":"string","default":""},     // blood pressure
      {"name": "drug", "type": "int", "default": ""},    // drug administered
]}
```

Também supõe que a chave seja modelada da seguinte maneira:

```
/u/<sex (M/F)>/<patient ID>
```

e preencha o armazenamento de dados usando estes comandos:

```
put -key /u/F/1 -value
  "{\"age\":\"23\",\"bp\":\"HIGH\",\"drug\":\"drugY\"}" -json avro.DrugSchema
put -key /u/M/2 -value
  "{\"age\":\"47\",\"bp\":\"LOW\",\"drug\":\"drugC\"}" -json avro.DrugSchema
put -key /u/M/3 -value
  "{\"age\":\"47\",\"bp\":\"LOW\",\"drug\":\"drugC\"}" -json avro.DrugSchema
put -key /u/F/4 -value
  "{\"age\":\"28\",\"bp\":\"NORMAL\",\"drug\":\"drugX\"}" -json avro.DrugSchema
put -key /u/F/5 -value
  "{\"age\":\"61\",\"bp\":\"LOW\",\"drug\":\"drugY\"}" -json avro.DrugSchema
```

Para preservar as informações sobre o sexo e o ID do usuário das chaves maiores, a tabela deve ser criada com um parâmetro adicional SERDEPROPERTIES **kv.avro.json.keys.mapping.column**. O valor do parâmetro deve ser o nome da coluna Hive do tipo map<string,string>. As chaves no mapa serão os nomes das chaves de registro especificadas nas propriedades **kv.*.keys.mapping** e os valores serão os valores da chave real. O DDL de criação de tabela é mostrado a seguir:

```
CREATE EXTERNAL TABLE oracle_user(keys map<string,string>, age string, bp string, drug string)
  STORED BY 'com.ibm.spss.hcatalog.OracleKVStorageHandler'
  WITH SERDEPROPERTIES ("kv.major.keys.mapping" = "DrugSchema,sex,patientid",
      "kv.parent.key" = "/u",
      "kv.avro.json.key" = "value",
      "kv.avro.json.keys.mapping.column" = "keys")
  TBLPROPERTIES ("kv.host.port" = "<hostname>:5000", "kv.name" = "kvstore");
```

Executar select * from oracle_user produzirá os resultados a seguir:

```
select * from
       oracle_user; {"user":"u","gender":"m"," userid":"125"} joe smith 77 13
      {"user":"u","gender":"m"," userid":"129"} jeff smith 67 27
      {"user":"u","gender":"m"," userid":"127"} jim smith 78 11
      {"user":"u","gender":"f"," userid":"131"} jen schmitt 70 20
      {"user":"u","gender":"m"," userid":"130"} jed schmidt 60 31
      {"user":"u","gender":"f"," userid":"128"} jan smythe 79 10
      {"user":"u","gender":"f"," userid":"126"} jess smith 76 12
```

A tabela oracle_user pode ser usada no console do Analytic Server para criar uma origem de dados do Oracle NoSQL. As chaves de sexo e ID do usuário, bem como os nomes da coluna do esquema Avro, podem ser usadas para definir os campos correspondentes para a origem de dados.

Varreduras de intervalo

O Analytic Server suporta varreduras de intervalo com base no prefixo pai para as chaves maiores, bem como os subintervalos para restringir ainda mais o intervalo na chave-pai.

A chave-pai especifica o prefixo para os pares de valores de chave "filho" a serem retornados. Um prefixo vazio resulta na busca de todas as chaves no armazenamento. Se o prefixo não estiver vazio, o caminho de chave maior deverá ser um caminho parcial e o caminho de chave menor deverá estar vazio. A chave pai é armazenada como um atributo de origem de dados **com.ibm.spss.ae.hcatalog.range.parent**.

O subintervalo restringe ainda mais o intervalo na chave-pai para os componentes do caminho maior no subintervalo. A chave de início de subintervalo é armazenada como **com.ibm.spss.ae.hcatalog.range.start** e a chave de término de subintervalo é armazenada como **com.ibm.spss.ae.hcatalog.range.end**. A chave de início deve ser lexicograficamente menor ou igual à chave de término. Os parâmetros de subintervalo são opcionais.

Origens de dados XML

O Analytic Server fornece suporte para os dados XML por meio do HCatalog.

Exemplo

1. Mapeie o esquema XML para os tipos de dados Hive por meio do Data Definition Language (DDL) do Hive, de acordo com as regras a seguir.

```
CREATE [EXTERNAL] TABLE <table_name> (<column_specifications>)
ROW FORMAT SERDE "com.ibm.spss.hive.serde2.xml.XmlSerDe"
WITH SERDEPROPERTIES (
    ["xml.processor.class"="<xml_processor_class_name>",]
    "column.xpath.<column_name>"="<xpath_query>",
     ...
    ["xml.map.specification.<element_name>"="<map_specification>"
     ...
    ]
)
STORED AS
    INPUTFORMAT "com.ibm.spss.hive.serde2.xml.XmlInputFormat"
    OUTPUTFORMAT "org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat"
[LOCATION "<data_location>"]
TBLPROPERTIES (
    "xmlinput.start"="<start_tag ",
    "xmlinput.end"="<end_tag>"
);
```

Nota: Se os arquivos XML estiverem compactados com a compactação Bz2, o `INPUTFORMAT` deverá ser configurado como `com.ibm.spss.hive.serde2.xml.SplittableXmlInputFormat`. Se eles estiverem compactados com a compactação CMX, deverá ser configurado como `com.ibm.spss.hive.serde2.xml.CmxXmlInputFormat`.

Por exemplo, o XML a seguir...

```
<records>
  <record customer_id="0000-JTALA">
    <demographics>
      <gender>F</gender>
      <agecat>1</agecat>
      <edcat>1</edcat>
      <jobcat>2</jobcat>
      <empcat>2</empcat>
      <retire>0</retire>
      <jobsat>1</jobsat>
      <marital>1</marital>
      <spousedcat>1</spousedcat>
      <residecat>4</residecat>
      <homeown>0</homeown>
      <hometype>2</hometype>
      <addresscat>2</addresscat>
    </demographics>
    <financial>
      <income>18</income>
      <creddebt>1.003392</creddebt>
```

```
      <othdebt>2.740608</othdebt>
      <default>0</default>
    </financial>
  </record>
</records>
```

...deve ser representado pelo DDL do Hive a seguir.

```
CREATE TABLE xml_bank(customer_id STRING, demographics map<string,string>, financial map<string,string>)
ROW FORMAT SERDE 'com.ibm.spss.hive.serde2.xml.XmlSerDe'
WITH SERDEPROPERTIES (
    "column.xpath.customer_id"="/record/@customer_id",
    "column.xpath.demographics"="/record/demographics/*",
    "column.xpath.financial"="/record/financial/*"
)
STORED AS
    INPUTFORMAT 'com.ibm.spss.hive.serde2.xml.XmlInputFormat'
    OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io.IgnoreKeyTextOutputFormat'
TBLPROPERTIES (
    "xmlinput.start"="<record customer",
    "xmlinput.end"="</record>"
);
```

Consulte "XML para mapeamentos de tipos de dados do Hive" para obter mais informações.

2. Crie uma origem de dados do Analytic Server com o tipo de conteúdo HCatalog no console do Analytic Server.

Limitações

- Somente a especificação XPath 1.0 é suportada atualmente.
- A parte local dos nomes qualificados para os elementos e atributos é usada ao manipular nomes do campo Hive. Os prefixos de namespace são ignorados.

XML para mapeamentos de tipos de dados do Hive: Os dados modelados no XML podem ser transformados para os tipos de dados do Hive usando as convenções documentadas a seguir.

Estruturas

O elemento XML pode ser mapeado diretamente para o tipo de estrutura do Hive, para que todos os atributos se tornem os membros de dados. O conteúdo do elemento se torna um membro adicional do tipo primitivo ou complexo.

dados XML

```
<result name="ID_DATUM">03.06.2009</result>
```

DDL do Hive e dados brutos

```
struct<name:string,result:string>
{"name":"ID_DATUM", "result":"0.3.06.2009"}
```

Matrizes

As sequências XML dos elementos podem ser representadas como matrizes do Hive do tipo primitivo ou complexo. O exemplo a seguir mostra como o usuário pode definir uma matriz de sequências usando o conteúdo do elemento <result> do XML.

dados XML

```
<result>03.06.2009</result>
<result>03.06.2010</result>
<result>03.06.2011</result>
```

DDL do Hive e dados brutos

```
result array<string>
{"result":["03.06.2009","03.06.2010",...]}
```

Mapas

O esquema XML não fornece suporte nativo para os mapas. Há três abordagens comuns para a modelagem de mapas no XML. Para acomodar as diferentes abordagens, usamos a sintaxe a seguir:

```
"xml.map.specification.<element_name>"="<key>-><value>"
```

em que

element_name
O nome do elemento XML a ser considerado como uma entrada de mapa

chave O nó XML da chave de entrada de mapa

valor O nó XML do valor da entrada de mapa

A especificação do mapa para o elemento XML fornecido deve ser definida na seção SERDEPROPERTIES no DDL de criação da tabela Hive. As chaves e valores podem ser definidos usando a sintaxe a seguir:

@attribute
A especificação @attribute permite ao usuário usar o valor do atributo como uma chave ou um valor do mapa.

elemento
O nome de elemento pode ser usado como uma chave ou um valor.

#conteúdo
O conteúdo do elemento pode ser usado como uma chave ou um valor. Visto que as chaves de mapa podem ser somente do tipo primitivo o conteúdo complexo será convertido para sequência.

As abordagens para representar mapas no XML e seus dados brutos e DDL do Hive, é conforme se segue.

Nome do elemento para conteúdo
O nome do elemento é usado como uma chave e o conteúdo como um valor. Essa é uma das técnicas comuns e é usada, por padrão, ao mapear o XML para os tipos de mapa Hive. A limitação evidente com essa abordagem é que a chave de mapa pode ser somente de sequência de tipos.

dados XML

```
<entry1>value1</entry1>
<entry2>value2</entry2>
<entry3>value3</entry3>
```

Mapeamento, DDL do Hive e dados brutos

Nesse caso, não é necessário especificar um mapeamento, pois o nome do elemento é usado como uma chave e o conteúdo como um valor, por padrão.

```
result map<string,string>
{"result":{"entry1": "value1", "entry2": "value2", "entry3": "value3"}}
```

Atributo para o Conteúdo de Elemento
Use um valor de atributo como uma chave e o conteúdo de elemento como um valor.

dados XML

```
<entry name="key1">value1</entry>
<entry name="key2">value2</entry>
<entry name="key3">value3</entry>
```

Mapeamento, DDL do Hive e dados brutos

```
"xml.map.specification.entry"="@name->#content"
result map<string,string>
{"result":{"key1": "value1", "key2": "value2", "key3": "value3"}}
```

Atributo para Atributo

dados XML

```
<entry name="key1" value="value1"/>
<entry name="key2" value="value2"/>
<entry name="key3" value="value3"/>
```

Mapeamento, DDL do Hive e dados brutos

```
"xml.map.specification.entry"="@name->@value"
result map<string,string>
{"result":{"key1": "value1", "key2": "value2", "key3": "value3"}}
```

Conteúdo complexo

O conteúdo complexo que está sendo usado como um tipo primitivo será convertido para uma sequência XML válida incluindo um elemento raiz denominado <string>. Considere o XML a seguir:

```
<dataset>
 <value>10</value>
 <value>20</value>
 <value>30</value>
</dataset>
```

A expressão XPath /dataset/* resultará em um número de nós XML <value> sendo retornado. Se o campo de destino for do tipo primitivo, a implementação transformará o resultado da consulta para o XML válido, incluindo o nó raiz <string>.

```
<string>
 <value>10</value>
 <value>20</value>
 <value>30</value>
</string>
```

Nota: A implementação não incluirá um elemento raiz <string>, se o resultado da consulta for um único elemento XML.

Conteúdo de texto

O conteúdo de texto somente espaços em branco de um elemento XML é ignorado.

Visualização e Metadados (Origem de Dados)

Clicar em **Visualização e Metadados** exibe uma amostra de registros e o modelo de dados para a origem de dados. Aqui você tem a chance de revisar as informações de metadados básicas.

Visualização
A guia Visualização mostra uma pequena amostra de registros e seus valores de campo.

Editar

A guia Editar exibe os metadados de campo básicos. Para origens de dados com o tipo de conteúdo Arquivos, o modelo de dados é gerado a partir de uma pequena amostra de registros e é possível editar manualmente os metadados de campo nesta guia. Para origens de dados com o tipo de conteúdo HCatalog, o modelo de dados é gerado com base nos Mapeamentos do Campo do HCatalog e não é possível editar o mapeamento de campo nesta guia.

Campo
Dê um clique duplo no nome do campo para editá-lo.

Medição
Esse é o nível de mediação, usado para descrever características dos dados em um determinado campo.

Papel Usado para informar aos nós de modelagem se os campos serão de Entrada (campos de

previsão) ou de Destino (campos previstos) para um processo de aprendizado de máquina. Ambos e Nenhum também são funções disponíveis, juntamente com Partição, o que indica que um campo usado para registros de partição em amostras separadas para treinamento, teste e validação. O valor Divisão especifica que modelos separados serão construídos para cada possível valor do campo. A frequência especifica que os valores de um campo devem ser usados como uma ponderação de frequência para cada registro. O ID de registro é usado para identificar um registro na saída.

Armazenamento
Armazenamento descreve a maneira como os dados são armazenados em um campo. por exemplo, um campo com valores de 1 e 0 armazena dados de número inteiro. Isso é diferente do nível de medição, que descreve o uso dos dados e não afeta o armazenamento. Por exemplo, você pode desejar configurar o nível de medição para um campo de número inteiro com valores de 1 e 0 para Sinalizador. Isso geralmente indica que 1 = True e 0 = False.

Valores
Mostra os valores individuais para os campos com medição categórica ou o intervalo de valores para campos com medição contínua.

Estrutura
Indica se os registros no campo contêm um valor único (Primitivos) ou uma lista de valores.

Espessura
Indica a espessura de uma lista; 0 é uma lista de primitivos, 1 é uma lista de listas, etc.

Varrer todos os valores de dados
Isso permite iniciar e cancelar a varredura dos valores de dados da origem de dados para determinar os valores de categoria e limites do intervalo. Se uma varredura estiver em andamento, clique no botão **Cancelar varredura de dados**. Varrer todos os valores de dados garante que os metadados estejam corretos, mas poderá levar algum tempo, se a origem de dados tiver muitos campos e registros.

Projetos são áreas de trabalho para armazenamento de entradas e acesso de saídas para tarefas. Eles fornecem a estrutura organizacional de nível superior para arquivos e pastas de contenção. Projetos podem ser compartilhados com usuários individuais e grupos.

Listagem de projeto

A página principal Projetos fornece uma lista de projetos dos quais o usuário atual é um membro.

- Clique em um nome do projeto para exibir seus detalhes e editar suas propriedades.
- Digite na área de procura para filtrar a listagem para exibir somente os projetos com a sequência de caracteres de procura em seu nome.
- Clique em **Novo** para criar um novo projeto com o nome especificado no diálogo **Incluir novo projeto**. Consulte “Nomeando regras” na página 28 para obter restrições nos nomes que é possível fornecer aos projetos.
- Clique em **Excluir** para remover o(s) projeto(s) selecionado(s). Essa ação remove o projeto e exclui todos os dados associados ao projeto do HDFS.
- Clique em **Atualizar** para atualizar a listagem.

Detalhes do projeto individual

A área de conteúdo é dividida nas seções recolhíveis **Detalhes**, **Compartilhamento**, **Arquivos** e **Versões**.

Detalhes

Nome Um campo de texto editável que mostra o nome do projeto.

Nome de exibição

Um campo de texto editável que mostra o nome do projeto, conforme exibido em outros aplicativos. Se estiver em branco, o Nome será usado como o nome de exibição.

Descrição

Um campo de texto editável para fornecer um texto explicativo sobre o projeto.

Versões a serem mantidas

Exclui automaticamente a versão mais antiga do projeto confirmado quando o número de versões excede o número especificado. O padrão é 25.

Nota: O processo de limpeza não é imediato, mas é executado no segundo plano a cada 20 minutos.

É público

Uma caixa de seleção que indica se alguém pode ver o projeto (marcada) ou se os usuários e grupos devem ser incluídos explicitamente como membros (desmarcada).

Clique em **Salvar** para manter o estado atual das configurações.

Compartilhamento

É possível compartilhar um projeto incluindo usuários e grupos como autores ou visualizadores.

- Digitando os filtros da caixa de texto nos usuários e grupos com a sequência de caracteres de procura em seu nome. Selecione o nível de compartilhamento e clique em **Incluir membro** para incluir na lista de membros.
  - Os autores são membros completos de um projeto e podem modificar o projeto, bem como as pastas e arquivos dentro dele. Esses usuários e membros desses grupos terão acesso de gravação (nó do Analytic Server Export) a este projeto ao tentarem se conectar ao Analytic Server através de IBM® SPSS Modeler.
  - Os visualizadores podem ver as pastas e arquivos dentro de um projeto e definir as origens de dados nos objetos dentro de um projeto, mas não podem modificar o projeto.
- Para remover um autor, selecione um usuário ou grupo na lista Autor e clique em **Remover membro**.

Nota: Os administradores têm acesso de leitura e gravação a cada projeto, independentemente se eles estão listados especificamente como um membro.

Nota: As mudanças feitas em Compartilhamento são aplicadas imediata e automaticamente.

Arquivos

Área de janela da estrutura do projeto

A área de janela direita mostra a estrutura de projeto/pasta para o projeto atualmente selecionado. É possível navegar na estrutura de pasta, mas ela não é editável, exceto por meio de botões.

- Clique em **Fazer o download do arquivo para o sistema de arquivos local** para fazer o download de um arquivo selecionado para o sistema de arquivos local.
- Clique em **Excluir o(s) projeto(s) selecionado(s)** para remover o arquivo/pasta selecionado.

Visualizador de Arquivos

Mostra a estrutura de pasta para o projeto atual. A estrutura de pasta é editável somente nos projetos definidos. Isto é, não é possível incluir arquivos, criar pastas ou excluir itens no nível raiz do modo **Projetos**. Para criar/excluir um projeto, retorne para a listagem Projeto.

- Clique em **Fazer upload do arquivo para HDFS** para fazer upload de um arquivo para o projeto/subpasta atual.

- Clique em **Criar uma nova pasta** para criar uma nova pasta na pasta atual, com o nome especificado no diálogo **Nome da nova pasta**.
- Clique em **Fazer download do arquivo no sistema de arquivos local** para fazer download dos arquivos selecionados no sistema de arquivos local.
- Clique em **Excluir o(s) arquivo(s) selecionado(s)** para remover os arquivos/pastas selecionados.

Versões

Projetos têm sua versão baseada nas mudanças no conteúdo do arquivo e pasta. As mudanças nos atributos de um projeto, como descrição, se é público e com quem é compartilhado, não requerem uma nova versão. A inclusão, modificação ou exclusão de arquivos ou pastas não requer uma nova versão.

Tabela de controle de versão do projeto

A tabela exibe as versões do projeto existentes, suas datas de criação e confirmação, os usuários responsáveis por cada versão e a versão pai. A versão pai é aquela na qual a versão selecionada é baseada.

- Clique em **Bloquear** para fazer mudanças nos conteúdos da versão do projeto selecionada.
- Clique em **Confirmar** para salvar todas as mudanças feitas em um projeto e fazer com que essa versão seja o estado visível atual do projeto.
- Clique em **Descartar** para descartar todas as mudanças feitas em um projeto bloqueado e retornar o estado visível do projeto para a versão confirmada mais recentemente.
- Clique em **Excluir** para remover a versão selecionada.

Os administradores podem gerenciar as regras de usuários e grupos por meio da página Usuários.

A área de conteúdo é dividida nas seções recolhíveis **Detalhes** e **Principal**.

Detalhes

Nome Um campo de texto não editável que exibe o nome do locatário.

Descrição
Um campo de texto editável que permite fornecer o texto explicativo sobre o locatário.

URL Essa é a URL a ser fornecida aos usuários para efetuar login no locatário através do console do Analytic Server.

Diretores

Diretores são usuários e grupos que são extraídos do provedor de segurança instalado durante a configuração. É possível alterar a regra de diretores para serem Administradores ou Usuários.

Métricas

Permite a configuração de limites de recursos para um locatário. Relata o espaço em disco atualmente usado pelo locatário.

- É possível configurar uma cota máxima de espaço em disco para o locatário; quando esse limite for atingido, nenhum dado poderá ser gravado no disco nesse locatário até que o espaço em disco suficiente seja limpo para trazer o uso de espaço em disco do locatário abaixo da cota.
- É possível configurar um nível de aviso de espaço em disco para o locatário; quando essa cota for excedida, nenhuma tarefa analítica poderá ser enviada para os principais nesse locatário até que o espaço em disco suficiente seja limpo para trazer o uso de espaço em disco do locatário abaixo da cota.

- É possível configurar um número máximo de tarefas paralelas que podem ser executadas em um momento único nesse locatário; quando essa cota for excedida, nenhuma tarefa analítica poderá ser enviada pelos principais nesse locatário até que uma tarefa de execução seja concluída.
- É possível configurar o número máximo de campos que uma origem de dados pode possuir. O limite será verificado quando uma origem de dados for criada ou atualizada.
- É possível configurar o número máximo de registros que uma origem de dados pode possuir. O limite será verificado quando uma origem de dados for criada ou atualizada; por exemplo, ao incluir um novo arquivo ou alterar as configurações para um arquivo.
- É possível configurar o tamanho máximo do arquivo em megabytes. O limite será verificado quando um arquivo for transferido por upload.

Nomeando regras

Para tudo que pode ser fornecido um nome exclusivo no Analytic Server, como origens de dados e projetos, as regras a seguir são aplicadas a esses nomes.

- Os nomes devem ser exclusivos nos objetos do mesmo tipo. Por exemplo, duas origens de dados não podem ser nomeadas insuranceClaims, mas uma origem de dados e um projeto pode cada um ser nomeado insuranceClaims.
- Os nomes fazem distinção entre maiúsculas e minúsculas. Por exemplo, insuranceClaims e InsuranceClaims são considerados nomes exclusivos.
- Os nomes ignoram espaço em brando à esquerda e à direita.
- Os caracteres a seguir são inválidos nos nomes.
  `~, #, %, &, *, {, }, \\, :, <, >, ?, /, |, ", \t, \r, \n`

Capítulo 3. Integração do SPSS Modeler

O SPSS Modeler é um ambiente de trabalho de mineração de dados que possui uma abordagem visual para análise. Cada ação distinta em uma tarefa, desde o acesso aos dados até a mesclagem de registros para gravação de um novo arquivo ou construção de um modelo, é representada por um nó na tela. Nós unimos todas essas ações para formar um fluxo analítico.

Para construir um fluxo do SPSS Modeler que possa ser executado com relação uma origem de dados do Analytic Server, inicie com um nó do Analytic Server Source. O SPSS Modeler empurrará o fluxo de volta ao Analytic Server, tanto quanto possível, em seguida, se necessário, puxará um subconjunto dos registros para concluir a execução do fluxo "localmente" no servidor SPSS Modeler. É possível configurar o número máximo de registros que o SPSS Modeler fará download nas propriedades de fluxo do Analytic Server.

Se a análise terminar com registros gravados de volta no HDFS, conclua o fluxo com um nó do Analytic Server Export.

Consulte a documentação do SPSS Modeler para obter detalhes sobre esses nós.

Nós Suportados

Muitos nós do SPSS Modeler são suportados para execução no HDFS, mas pode haver algumas diferenças na execução de certos nós, e alguns não são atualmente suportados. Esse tópico detalha o nível atual de suporte.

Geral

- Alguns caracteres que são normalmente aceitos dentro de um nome de campo de Modelador entre aspas não serão aceitos pelo Analytic Server.
- Para um fluxo do Modeler a ser executado no Analytic Server, ele deve iniciar com um ou mais nós do Analytic Server Source e terminar em um único nó de modelagem ou nó do Analytic Server Export.
- É recomendado que você configure o armazenamento de variáveis respostas contínuas como número real em vez de número inteiro. Modelos de pontuação sempre gravam valores de número real para os arquivos de dados de saída para variáveis respostas contínuas, enquanto o modelo de dados de saída para as pontuações segue o armazenamento do destino. Assim, se uma variável resposta contínua tiver um armazenamento de número inteiro, haverá uma incompatibilidade entre os valores gravados e o modelo de dados para as pontuações e essa incompatibilidade causará erros quando for tentada a leitura dos dados de pontuação.

Fonte

- Um fluxo que começa com qualquer coisa que não seja um nó de origem do Analytic Server será executado localmente.

Operações de Registro

Todas as Operações de registro são suportadas, com a exceção dos nós TS do fluxo e Caixas de espaço-tempo. Notas adicionais sobre acompanhamento da funcionalidade de nó suportado.

Selecionar

- Suporta o mesmo conjunto de funções suportadas pelo Derivar nó.

Amostra

- Amostragem de nível de bloqueio não é suportada.
- Métodos de Amostragem complexos não são suportados.

Agregado

- Chaves contíguas não são suportadas. Se você estiver reutilizando um fluxo existente configurado para classificar os dados e, em seguida, usar essa configuração no nó Agregado, altere o fluxo para remover o nó Classificação.
- As estatísticas de ordem (Mediana, primeiro Quartil, terceiro Quartil) são calculadas aproximadamente e suportadas através da guia Otimização.

Classificar

- A guia Otimização não é suportada.

Em um ambiente distribuído, há um número limitado de operações que preservam a ordem de registro estabelecida pelo nó Classificação.

- Uma classificação seguida por um nó Exportação produz uma origem de dados classificados.
- Uma classificação seguida por um nó Amostra com a amostragem **Primeiro** registro retorna os primeiros registros *N*.
- Uma classificação seguida por um nó de modelagem que tem o objetivo **Otimizar para conjuntos de dados muito grandes** (Rede Neural, Linear, Árvore C&R, Quest ou CHAID) é um padrão útil para remanejar aleatoriamente os registros, classificando por uma chave de número aleatório derivado para evitar viés que podem ser introduzidas no algoritmo de construção de modelo, se os registros originais forem ordenados.

Em geral, deve-se colocar um nó Classificar o mais próximo possível das operações que precisam de registros classificados.

Mesclar

- Mesclagem por Ordem não é suportada.
- A guia Otimização não é suportada.
- Atualmente não é suportado colocar um nó de Amostra ou nugget de modelo entre um nó de Origem do Analytic Server e um nó de Mesclagem. Normalmente é possível especificar um nó de Seleção para substituir a funcionalidade do nó de Amostra.
- O Analytic Server não se associa a chaves de sequência vazia; isto é, se uma das chaves pelas quais você estiver mesclando contiver sequências vazias, então quaisquer registros que contiverem a sequência vazia serão descartados da saída mesclada.
- Operações de mesclagem são relativamente lentas. Se você tiver espaço disponível no HDFS, pode ser muito mais rápida mesclar suas origens de dados uma vez e usar a origem mesclada nos seguintes fluxos do que mesclar as origens de dados em cada fluxo.

Transformação de R

A sintaxe R no nó deve consistir em operações de registro em um momento.

Operações de campo

Todas as operações Campo são suportadas, com a exceção dos nós Transpor, Intervalos de Tempo e Histórico. Notas adicionais sobre acompanhamento da funcionalidade de nó suportado.

Preparação de dados automática

- Treinamento de nó não é suportado. Aplicar as transformações em um nó de Preparação de Dados Automática treinado para novos dados é suportado.

Tipo

- A coluna Verificação não é suportada.
- A guia Formato não é suportada.

Derivação

- Todas as funções de Derivação são suportadas, com exceção de funções de sequência.

- Campos de divisão não podem ser derivados no mesmo fluxo que os use como divisões; você precisará criar dois fluxos; um que derive o campo de divisão e um que use o campo como divisões.
- Um campo sinalizador não pode ser usado sozinho em uma comparação; isto é, if (flagField) then ... endif causará um erro; a solução alternativa é usar if (flagField=trueValue) then ... endif
- Isto é recomendado ao usar o operador ** para especificar o expoente como um número real, tal como x**2.0, em vez de f x**2, para corresponder resultados no Modelador

Preenchedor

- Suporta o mesmo conjunto de funções suportadas pelo Derivar nó.

Categorização

A funcionalidade a seguir não é suportada.

- Ótima categorização
- Ranqueamentos
- Ladrilhos -> Agrupamento lado a lado: Soma de valores
- Ladrilhos -> Ligações: Manter em atual e designar aleatoriamente
- Ladrilhos ->N customizado: Valores acima de 100 e qualquer valor N em que 100 % N não é igual a zero.

Análise RFM

- A opção Manter em atual para ligações de manipulação não é suportado. Pontuações de recência, frequência e monetárias do RFM nem sempre corresponderão àquelas calculadas pelo Modelador dos mesmos dados. Os intervalos de pontuação serão os mesmos, mas as designações de pontuação (número de categoria) podem diferir em um.

Gráficos

Todos os nós de Gráfico são suportados.

Modelagem

Os nós Modelagem a seguir são suportados: Linear, Rede Neural, C&RT, Chaid, Quest, TCM, TwoStep-AS, STP e Regras de Associação. Notas adicionais sobre a funcionalidade desses nós a seguir.

Linear Ao construir modelos em big data, normalmente você desejará alterar o objetivo para conjuntos de dados muito grandes ou especificar as divisões.

- Treinamento contínuo de modelos PSM existentes não é suportado.
- O objetivo da construção do modelo Padrão apenas será recomendado se os campos de divisão forem definidos de modo que número de registros em cada divisão não seja muito alto, sendo que a definição de "muito alto" depende do poder dos nós individuais no cluster do Hadoop. Por contraste, também é necessário ser cuidadoso para assegurar que as divisões não sejam definidas em partes muito pequenas de maneira que haja muito poucos registros para construir um modelo.
- O objetivo de Boosting não é suportado.
- O objetivo de Bagging não é suportado.
- O objetivo dos conjuntos de dados muito grandes não é recomendado quando há poucos registros; frequentemente ele não construirá um modelo ou construirá um modelo degradado.
- Preparação de Dados Automática não é suportada. Isso pode causar problemas ao tentar construir um modelo sobre dados com muitos valores ausentes; normalmente esse seriam inseridos como parte da preparação de dados automática. Uma solução

alternativa seria usar um modelo de árvore ou uma rede neural com a configuração Avançado para inserir valores ausentes selecionados.

- A estatística de previsão não é calculada para modelos de divisão.

Rede Neural

Ao construir modelos em big data, normalmente você desejará alterar o objetivo para conjuntos de dados muito grandes ou especificar as divisões.

- Treinamento contínuo de padrão existente ou modelos PSM não é suportado.
- O objetivo da construção do modelo Padrão apenas será recomendado se os campos de divisão forem definidos de modo que número de registros em cada divisão não seja muito alto, sendo que a definição de "muito alto" depende do poder dos nós individuais no cluster do Hadoop. Por contraste, também é necessário ser cuidadoso para assegurar que as divisões não sejam definidas em partes muito pequenas de maneira que haja muito poucos registros para construir um modelo.
- O objetivo de Boosting não é suportado.
- O objetivo de Bagging não é suportado.
- O objetivo dos conjuntos de dados muito grandes não é recomendado quando há poucos registros; frequentemente ele não construirá um modelo ou construirá um modelo degradado.
- Quando houver muitos valores ausentes nos dados, use a configuração Avançado para inserir valores ausentes.
- A estatística de previsão não é calculada para modelos de divisão.

Árvore C&R, CHAID e Quest

Ao construir modelos em big data, normalmente você desejará alterar o objetivo para conjuntos de dados muito grandes ou especificar as divisões.

- Treinamento contínuo de modelos PSM existentes não é suportado.
- O objetivo da construção do modelo Padrão apenas será recomendado se os campos de divisão forem definidos de modo que número de registros em cada divisão não seja muito alto, sendo que a definição de "muito alto" depende do poder dos nós individuais no cluster do Hadoop. Por contraste, também é necessário ser cuidadoso para assegurar que as divisões não sejam definidas em partes muito pequenas de maneira que haja muito poucos registros para construir um modelo.
- O objetivo de Boosting não é suportado.
- O objetivo de Bagging não é suportado.
- O objetivo dos conjuntos de dados muito grandes não é recomendado quando há poucos registros; frequentemente ele não construirá um modelo ou construirá um modelo degradado.
- Sessões interativas não são suportadas.
- A estatística de previsão não é calculada para modelos de divisão.

Pontuação do modelo

Todos os modelos suportados para modelagem também são suportados para pontuação. Além disso, nuggets do modelo construído localmente para os nós a seguir são suportados para escoragem: C&RT, Quest, CHAID, Linear e Rede Neural (independentemente se o modelo é padrão, empacotado impulsionado ou para conjuntos de dados muito grandes), Regressão, C5.0, Logística, Genlin, GLMM, Cox, SVM, Rede de Bayes, TwoStep, KNN, Lista de Decisão, Discriminante, Autoaprendizado, Detecção de anomalia, Apriori, Carma, K-Médias, Kohonen, R e Mineração de texto.

- Nenhuma propensão ajustada ou bruta será pontuada. Como uma solução alternativa é possível obter o mesmo efeito calculando manualmente a propensão bruta usando um nó Derivar com as seguintes expressões: if 'predicted-value' == 'value-of-interest' then 'prob-of-that-value' else 1-'prob-of-that-value' endif

- Ao pontuar um modelo, o Analytic Server não verifica se todos os campos usados no modelo estão presentes no conjuntos de dados, assim, certifique-se de que isso é verdadeiro antes de executar no Analytic Server

R A sintaxe R na nugget deverá consistir em operações de registro em um momento.

Out Os nós de Matriz, Análise, Auditoria de Dados, Transformação, Estatísticas e Médias são suportados.

O nó Tabela é suportado escrevendo uma origem de dados do Analytic Server temporário que contém os resultados de operações de envio de dados. O nó Tabela, em seguida, pagina através do conteúdo dessa origem de dados.

Exportar

Um fluxo pode iniciar com um nó de origem do Analytic Server e terminar com um nó de exportação diferente do nó de exportação do Analytic Server, mas os dados se movimentarão do HDFS para o SPSS Modeler Server, e finalmente para a localização de exportação.

Avisos

Estas informações foram desenvolvidas para produtos e serviços oferecidos nos Estados Unidos.

É possível que a IBM não ofereça os produtos, serviços ou recursos discutidos nesta publicação em outros países. Consulte um representante IBM local para obter informações sobre produtos e serviços disponíveis atualmente em sua área. Qualquer referência a produtos, programas ou serviços IBM não significa que apenas produtos, programas ou serviços IBM possam ser utilizados. Qualquer produto, programa ou serviço funcionalmente equivalente, que não infrinja nenhum direito de propriedade intelectual IBM poderá ser utilizado em substituição. Entretanto, a avaliação e verificação da operação de qualquer produto, programa ou serviço não IBM são de responsabilidade do Cliente.

A IBM pode ter patentes ou solicitações de patentes pendentes relativas a assuntos tratados nesta publicação. O fornecimento desta publicação não lhe concede direito algum sobre tais patentes. Consultas sobre licença devem ser enviados, por escrito, para:

Gerência de Relações Comerciais e Industriais da IBM
Brasil
Av. Pasteur, 138-146
Botafogo
Rio de Janeiro, RJ
CEP 22290-240

Para consultas sobre licença relacionados a informações de DBCS (Conjunto de Caracteres de Byte Duplo), entre em contato com o Departamento de Propriedade Intelectual da IBM em seu país ou envie consultas sobre licença, por escrito, para:

IBM World Trade Asia Corporation
Licensing
2-31 Roppongi 3-chome, Minato-ku
Tokyo 106
Japan

O parágrafo a seguir não se aplica a nenhum país em que tais disposições não estejam de acordo com a legislação local: A INTERNATIONAL BUSINESS MACHINES CORPORATION FORNECE ESTA PUBLICAÇÃO “NO ESTADO EM QUE SE ENCONTRA”, SEM GARANTIA DE NENHUM TIPO, SEJA EXPRESSA OU IMPLÍCITA, INCLUINDO, MAS A ELAS NÃO SE LIMITANDO, AS GARANTIAS IMPLÍCITAS DE NÃO INFRAÇÃO, COMERCIALIZAÇÃO OU ADEQUAÇÃO A UM DETERMINADO PROPÓSITO. Alguns países não permitem a exclusão de garantias expressas ou implícitas em certas transações; portanto, essa disposição pode não se aplicar ao Cliente.

Essas informações podem conter imprecisões técnicas ou erros tipográficos. São feitas alterações periódicas nas informações aqui contidas; tais alterações serão incorporadas em futuras edições desta publicação. A IBM pode, a qualquer momento, aperfeiçoar e/ou alterar os produtos e/ou programas descritos nesta publicação, sem aviso prévio.

Referências nestas informações a Web sites não IBM são fornecidas apenas por conveniência e não representam de forma alguma um endosso a esses Web sites. Os materiais contidos nesses Web sites não fazem parte dos materiais desse produto IBM e a utilização desses Web sites é de inteira responsabilidade do Cliente.

A IBM pode utilizar ou distribuir as informações fornecidas da forma que julgar apropriada sem incorrer em qualquer obrigação para com o Cliente.

Licenciados deste programa que desejam obter informações sobre este assunto com objetivo de permitir: (i) a troca de informações entre programas criados independentemente e outros programas (incluindo este) e (ii) a utilização mútua das informações trocadas, devem entrar em contato com:

Gerência de Relações Comerciais e Industriais da IBM Brasil
Av. Pasteur, 138-146
Botafogo
Rio de Janeiro, RJ
CEP 22290-240

Tais informações podem estar disponíveis, sujeitas a termos e condições apropriadas, incluindo em alguns casos o pagamento de uma taxa.

O programa licenciado descrito nesta publicação e todo o material licenciado disponível são fornecidos pela IBM sob os termos do Contrato com o Cliente IBM, do Contrato Internacional de Licença do Programa IBM ou de qualquer outro contrato equivalente.

Todos os dados de desempenho aqui contidos foram determinados em um ambiente controlado. Sendo assim, os resultados obtidos em outros ambientes operacionais podem variar significativamente. Algumas medidas podem ter sido tomadas em sistemas em nível de desenvolvimento e não há garantia de que estas medidas serão iguais em sistemas geralmente disponíveis. Além disso, algumas medidas podem ter sido estimadas por extrapolação. Os resultados reais podem variar. Os usuários deste documento devem verificar os dados aplicáveis para seu ambiente específico.

As informações relativas a produtos não IBM foram obtidas junto aos fornecedores dos respectivos produtos, de seus anúncios publicados ou de outras fontes disponíveis publicamente. A IBM não testou estes produtos e não pode confirmar a precisão de seu desempenho, compatibilidade nem qualquer outra reivindicação relacionada a produtos não IBM. Dúvidas sobre os recursos de produtos não IBM devem ser encaminhadas diretamente a seus fornecedores.

Todas as declarações relacionadas aos objetivos e intenções futuras da IBM estão sujeitas a alterações ou cancelamento sem aviso prévio e representam apenas metas e objetivos.

Todos os preços IBM mostrados são preços de varejo sugeridos pela IBM, são atuais e estão sujeitos a alteração sem aviso prévio. Os preços do revendedor podem variar.

Estas informações foram projetadas apenas com o propósito de planejamento. As informações aqui contidas estão sujeitas a alterações antes que os produtos descritos estejam disponíveis.

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com nomes e endereços utilizados por uma empresa real é mera coincidência.

Cada cópia ou parte destes programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

Estas informações contêm exemplos de dados e relatórios utilizados nas operações diárias de negócios. Para ilustrá-los da forma mais completa possível, os exemplos podem incluir nomes de indivíduos, empresas, marcas e produtos. Todos estes nomes são fictícios e qualquer semelhança com nomes e endereços utilizados por uma empresa real é mera coincidência.

Cada cópia ou parte destes programas de amostra ou qualquer trabalho derivado deve incluir um aviso de copyright com os dizeres:

© (nome da empresa) (ano). Partes deste código são derivadas dos Programas de Amostra da IBM Corp.

© Copyright IBM Corp. _insira o ano ou os anos_. Todos os direitos reservados.

Se estas informações estiverem sendo exibidas em cópia eletrônica, as fotografias e ilustrações coloridas podem não aparecer.

Marcas Registradas

A IBM, o logotipo IBM e ibm.com são marcas comerciais ou marcas registradas da International Business Machines Corp., registradas em muitas jurisdições em todo mundo. Outros nomes de produtos e serviços podem ser marcas comerciais da IBM ou de outras empresas. Uma lista atual de marcas comerciais da IBM está disponível na Web em "Copyright and trademark information" em www.ibm.com/legal/copytrade.shtml.

Adobe, o logotipo Adobe, PostScript e o logotipo PostScript são marcas comerciais ou marcas registradas da Adobe Systems Incorporated nos Estados Unidos, e/ou em outros países.

IT Infrastructure Library é uma marca registrada da Agência Central de Computação e Telecomunicações, que agora é parte do Departamento de Comércio do Governo.

Intel, logotipo Intel, Intel Inside, logotipo Intel Inside, Intel Centrino, logotipo Intel Centrino, Celeron, Intel Xeon, Intel SpeedStep, Itanium e Pentium são marcas comerciais ou marcas registradas da Intel Corporation ou de suas subsidiárias nos Estados Unidos e em outros países.

Linux é uma marca registrada de Linus Torvalds nos Estados Unidos e/ou em outros países.

Microsoft, Windows, Windows NT e o logotipo Windows são marcas comerciais da Microsoft Corporation nos Estados Unidos e/ou em outros países.

ITIL é uma marca registrada e uma marca registrada da comunidade do The Minister for the Cabinet Office e está registrada no U.S. Patent and Trademark Office.

UNIX é marca registrada do The Open Group nos Estados Unidos e/ou em outros países.

Java e todas as marcas comerciais e logotipos baseados em Java são marcas comerciais ou marcas registradas da Oracle e/ou de suas afiliadas.

Cell Broadband Engine é uma marca comercial da Sony Computer Entertainment, Inc. nos Estados Unidos e/ou em outros países e é utilizada sob licença a partir deste ponto.

Linear Tape-Open, LTO, o logotipo LTO, Ultrium e o logotipo Ultrium são marcas comerciais da HP, IBM Corp. e Quantum nos Estados Unidos e em outros países.

Impresso no Brasil